Classifying breast cancer subtypes based on aCGH profile

Emmanouil Bouzetos[1,2], Liliána Kádár[1,2], Molin Zheng[1,2], and Robert Zomerdijk[1,2]

[1] Vrije Universiteit Amsterdam, The Netherlands
[2] University of Amsterdam, The Netherlands

Abstract. Breast cancer is a heterogeneous diseases caused by DNA copy number alterations (CNAs). Multiple breast cancer subtypes can be identified based on histological or molecular characteristic. Knowing the subtypes provide us with prognostics information and also can guide therapeutic decisions. aCGH is an automatic, sensitive, high-resolution method to identify CNAs profile. Using such profile to classify breast cancer into subtypes would allow us to omit the cell culturing step in current subtype classification methods. In this project we used aCGH data to classify breast cancer into three subtypes based on hormone receptors. We have compared four classification models and found that Random Forest had the highest performance with 0.920 accuracy. We also found a promising biomarker regions on chromosome 17 (start: 35076296, end: 3528208). These findings can help to make breast cancer subtype classification easier, and also to understand what are the key DNA aberrations events that lead to tumorigenesis.

1 Introduction

Breast cancer is a heterogeneous disease that arises from the epithelial cells of the breast [12][14]. It can be classified into several histological and molecular subtypes, but the most widely accepted subtypes are defined based on the immunochemical quantitation of three hormone receptors' expression levels: estrogen (ER), progesterone (PR), and human epidermal growth factor (HER2). As a result, we can identify four subtypes: estrogen receptor positive (ER+), progesterone receptor positive (PR+), human epidermal growth factor receptor positive (HER2+), and triple-negative (TNBC) [12].

Knowing the subtype of breast cancer can help to gain prognostic information, or predict therapy responsiveness. For example, a higher expression of PR is associated with a better survival rate and recurrence time, while a lower expression implies a more aggressive course of disease [12]. However, subclassification is not the only factor that needs to be considered regarding prognosis, other factors such as age, tumour size, lymph node status, comorbidity and adjuvant therapy also need to be considered [10].

As chromosomal aberrations are frequently involved in cancer development, DNA copy number alterations(CNAs) can also be used as prognostic factors. Especially for breast cancer, which is a disease with high levels of chromosomal instability [1]. Molecular profiling can identify clinically relevant features of the tumour, such as subtypes luminal A and B were discovered. While both are estrogen receptor positive, the latter usually has a worse outcome [2]. It is hoped that CNAs can be used as prognostic and predictive factors for breast cancer.

Array-based Comparative Genomic Hybridization

Array-based comparative genomic hybridization (array-CGH or aCGH) is a reliable, sensitive and highly-automated method that profiles CNAs genome-wide at high-resolution [1]. It uses microarrays, that consists of thousands of genomic target or probe [9]. In order to detect chromosomal aberration the genome needs to be split into continuous segments. During calling the different segmentation states are classified as a loss, normal, gain or amplification [15]. However, genomic segments can form so-called regions, which are a series of neighbouring clones on the chromosomes whose aCGH-signature is the same. Using such regions instead of keeping every single chromosomal segment reduces the dimensionality of the data.

Array-CGH has been used to classify breast cancer into BRCA1 and BRCA2 subtypes based on genetic profiles [6]. It might have the potential to differentiate and classify between other subtypes. For example, loss of material in chromosome 16q is an early event in the pathway to a good-prognosis EG+ breast cancer prognoses [6].

1.1 Aims

In this project, we investigate if CNAs can be used to differentiate between three breast cancer subtypes: HER2 positive (HER+), hormone receptor positive (HR+) which includes PR+ and ER+ subtypes, and triple-negative (TN).

We will train several machine learning models (Random Forest, PLS-DA, Neural Network and Long Short-Term Memory) on 100 samples to select the best model for breast cancer subtype classification, and we also try to identify a specific chromosomal region that could be used as a single biomarker for such classification.

2 Methods

2.1 Data Description

Our data set consisted of 100 classified aCGH samples (arrays) with 2834 variables. The variables represented regions of the chromosomes and their values

could be: 0 (normal), loss (-1), gain (1) or 2 (amplification). The samples were classified into three breast cancer subtypes: HER2 positive (HER2+), Hormone receptor positive (HR+), or Triple negative (TN). The three classes are represented equally, as there are 32 HER2+, 32 Triple Negative and 36 HR+ samples.

As the data set was pre-cleaned for us we did not have to handle missing, incorrect or extreme values. The only feature engineering we performed is we combined the chromosome, start and column into one column containing the number of the chromosome and the index number of the regions. In addition, we created a separate data set explaining the start and end points of each chromosomal region.

2.2 Nested Cross-validation

Due to the low number of samples we used nested cross-validation to optimize hyperparameters and evaluate the performance of the models while decreasing overfitting. The outer loop was a 3-fold cross-validation, used to evaluate the performance of the models - that resulted in approximately 30:70 test: train data ratio for each fold. The inner loop was used for hyperparameter optimization and it was a 10-fold cross-validation. Fig. 1 provides an overview of the validation scheme. A disadvantage of nested cross-validation is that it significantly increases the number of models that need to be evaluated.

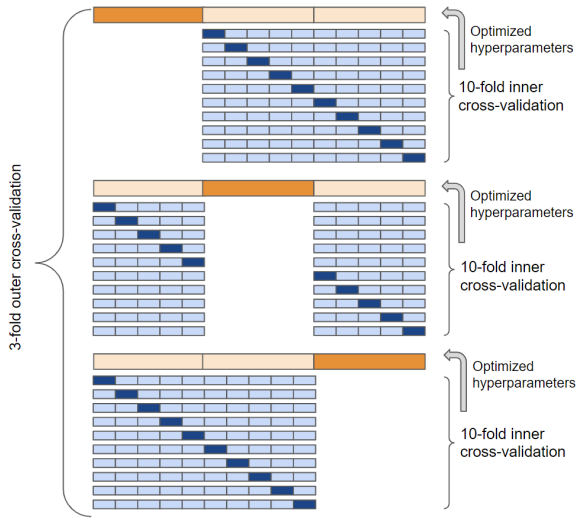

Fig. 1: *Nested cross-validation scheme: 3-fold outer and 10-fold inner cross-validation. Legend: dark orange - outer fold test set, light orange - outer fold training set, dark blue - inner fold test set, light blue - inner fold training set*

2.3 Feature Selection

We have used two different feature selection methods to identify the most important features (chromosomal regions) for breast cancer subtype prediction: Multivariate Methods with Unbiased Variable Selection in R package (MUVR) [13] by Carl Brunius and the inbuilt feature importance attribute of the scikit-learn Python package's Gradient Boosting Classifier (GBC).

MUVR selects the optimal number of features by performing recursive variable elimination in a repeated double cross-validation. It simultaneously identifies minimal-optimal and all-relevant variable sets for regression, classification and multilevel analyses [13]. We used its Random Forest modelling with the default setting for feature selection. This method not just rank the features but also selects the optimal number of features to achieve the lowest number of misclassification as illustrated by (Fig. 2). As the MUVR package found that 17 is the optimal number of features, we looked for the top 17 features when we applied the other feature selection method.

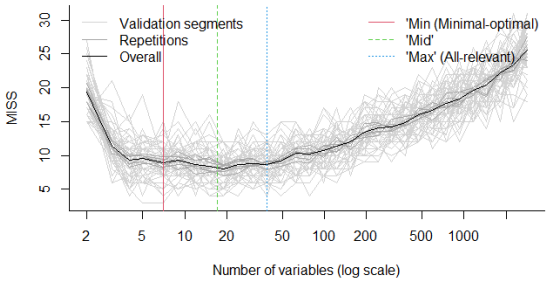

Fig. 2: *MUVR feature selection: plotting the number of misclassification as a function of number of variables. The green line shows that using 17 variable is the optimal in order to reach the lowest number of misclassification.*

The Gradient Boosting Classifier algorithm builds an additive model using n_classes_regression trees. It has a built-in feature_importances_ attribute that provides us with the importance of each feature regarding the Gradient Boosting Classifier, from which we can select the 17 most important ones. We used the scikit-learn Python package for this feature selection method.

There are two chromosomal regions that were selected by both feature selection methods:

- 17 (35076296:35282086)
- 22 (41307174:41912419)

Both methods selected regions from chromosome 5, 6, 17 and 22.

The exact list of what features were selected by the two feature selection methods can be found in the Appendix.

2.4 Baseline Model

Our baseline model is the Dummy Classifier of the Python scikit-learn package, and it was used with all available classification strategies - most frequent, prior,

stratified, uniform and constant. The constant strategy using HR+ as a constant value yielded the highest accuracy, 0.362, therefore we expect our models to perform better than that.

2.5 Model Optimization and Evaluation

Models were optimized in two different setups: using the GBC-selected features and the MUVR-selected features. That means two rounds of hyperparameter optimization with Grid Search, resulting in two mean accuracy scores that were used to select the best selected 17 features for the model, and then select the best model from the batch of models that were trained with that combination of selected features.

Once a best model and feature selection method were selected for a type of classifier, the selected best model was further evaluated in the 3-fold cross-validation using accuracy, precision, recall, F1 score and AUC as evaluation metrics for model comparison.

2.6 PLS-DA

The partial square discriminant analysis is a supervised classification algorithm. PLSDA is the combination of the dimension reduction and discriminant analysis, provide classification methods to high-dimensional data. This algorithm can be used for model prediction, but it can also be used for discriminate feature selection. However, this model also has some disadvantages, for example, the algorithm does not have the mechanism to help against overfitting.

We have performed a PLS-DA firstly on the original data. According to the score plot (which can be found in the appendix), the PLSDA model cannot distinguish the three breast cancer subtypes clearly. It can be seen that there are some overlaps between the three groups.

Then we used nested cross-validation to tune the ncomp and the orthoI parameters (the default ncomp is 2). Based on model accuracy, we still consider 2 ncomp as the optimal parameter, and the orthoI is 0.

We also looked at what 17 features the PLS-DA would identified as top features using variance importance (VIP), and compared it to the features selected by the MUVR and GBC methods. It also found the 17 (35076296:35282086) and 22 (41307174:41912419) regions that were found both MUVR and GBC, but in addition it has seven common regions with the MUVR, and two common regions with the GBC methods, supporting the ideas that these are the most important biomarker regions.

2.7 Random Forest Classification

Random Forest is an ensembled model, using multiple decision trees to perform classification. Decision tree is a greedy algorithm and prone to overfitting. In a Random Forest model the trees do not see neither all the feature nor all the samples which protect against overfitting. The final decision is usually made by majority voting. Optimizing hyperparameter can also help to reduce overfitting.

The previously described nested cross-validation was used to optimize the following hyperparemeters based on model accuracy: number of trees (10, 50, 100, 150, 200), maximum tree depth (5, 10, 15, 20), and the criteria (gini ,entropy) that is used for splitting. It was found that the optimal hyperparameter depends on the fragments of the data set that is involved in the training, and models that trained on the MUVR selected features yielded a higher accuracy. The most commonly selected model when the MUVR selected features were used for training was the Random Forest that was made out of 100 decision trees, with maximum depth 5, and used gini as the splitting criteria.

2.8 Neural network

Neural network (also known as artificial Neural Networks (aNNs)) is a machine learning algorithm, which at its heart is a deep learning algorithm. Neural networks are comprised of multiple layers. An input layer, one or more hidden layers and an output layer. However, neural networks usually performs poorly with low number of training data.

For this research we created a simple neural network using TensorFlow in python, and for the hyper parameter optimization we used the scikit-learn Python package. The optimization was done doing nested cross-validation and evaluating using GridSearch. The outer folds used a 3 split and the inner fold used a 10 split cross validation. For the hyper parametersh we optimized the following hyperparameters: hidden layer sizes(hls = (5,3), (5,2), (5,3,2), (5,4,3)), learning rate (lr = 0.01, 0.1, 0.2), activation function (af = relu, tanh) and alpha (a = 0.01, 0.1, 1).

Just as in the case of previously described models we found that the best model depends on which part of the data the model trains on. In addition, most of the tested models have reached the maximum iteration before the optimization has converged. To resolve this issue we tried to increase the number of maximum iteration but it did not help. Optimizing the model with the MUVR selected feature yielded higher accuracy so the final model was selected using the information from that hyperparameter optimization.

The final best neural network model uses relu as activation function, with a 0.2 learning rate and an alpha value of 0.01, and it has two hidden layers with 5 and 3 nodes.

2.9 Long Short-Term Memory

LSTM (Long Short-Term Memory) is a type of recurrent neural network (RNN) model commonly used for sequence-based tasks, such as time series prediction or natural language processing. LSTM models consist of memory cells that store information over time. These memory cells are responsible for selectively remembering or forgetting information based on the input data

and previous context. This ability to retain and utilize relevant information makes LSTM models particularly effective in capturing complex temporal patterns. Although our data are not temporal, typical neural networks assume independence across all observations in a series of data, thus we considered that would be interesting to explore the behaviour of a LSTM model in the use of our aCGH data set. The input array was transformed to the dimensions (100, 1, 17), giving each sample 1 feature with 17 different time-steps. To perform hyperparameter optimization, which could help preventing overfitting and improve performance, the nested cross-validation technique was selected. We tuned the number of layers (1,2,3), the number of neurons in each layer (10, 20, 30, 40, 50) and the type of the optimizer(adam, rmsprop, sgd). The LSTM model was trained and evaluated under two different setups. In the first setup,the model was trained with 17 features provided by MURV and in the second setup the subset of 17 features was selected using the GradientBoostingClassifier (GDB) technique. After thorough evaluation of the results of CV, our selected optimal LSTM model consisted of two layer with the first having 50 neurons, the second 10 and the optimizer Adam. Evaluating the the curves of Loss and validation loss 100 epocjs for training were choosen.

3 Results

3.1 Classifiers' Performance

We have built and optimized 4 classifiers (Random Forest (RF), PLS-DA, Neural Network (NN), and Long Short-Term Memory (LSTM)) using 17 features from 2 different feature selection methods to see which optimized model is the most suitable for breast cancer subtype classification. We have measured the models' performance using accuracy, precision, recall, F1 score and AUC. Table 1 shows the optimized models' performance and the feature selection method from which the top 17 features were used. As we used cross-validation, the numbers in the table represent the mean performance metrics.

Table 1: *Mean performance of the optimized classifiers when optimal top 17 features are used*

	PLS-DA	RF	NN	LSTM
Feature selection	MUVR	MUVR	MUVR	MUVR
Accuracy	0.905	0.920	0.870	0.890
Precision	0.912	0.932	0.872	0.908
Recall	0.890	0.920	0.870	0.878
F1	0.895	0.920	0.868	0.894
AUC	0.895	0.977	0.91	0.975

We can see that the Random Forest classifier has the highest performance values regardless of which metrics we are looking at, so we have selected the optimized Random Forest model as our classifier for the breast cancer subtype prediction. Its accuracy is 0.920, which is much higher accuracy than what we see from the dummy classifier that performed the best (0.36).

3.2 Detailed Evaluation of the Random Forest model

The optimized Random Forest model showed the best performance according to the five evaluation metrics we used. This is not surprising as it has been previously proven to be a good choice when the model uses genomics data for classification [3] [11] [5]. In addition, the MUVR feature selection method uses a Random Forest model when selecting the features, and while using only the MUVR-selected 17 features improved the performance of all model, it could have the highest impact on the Random Forest model because of it.

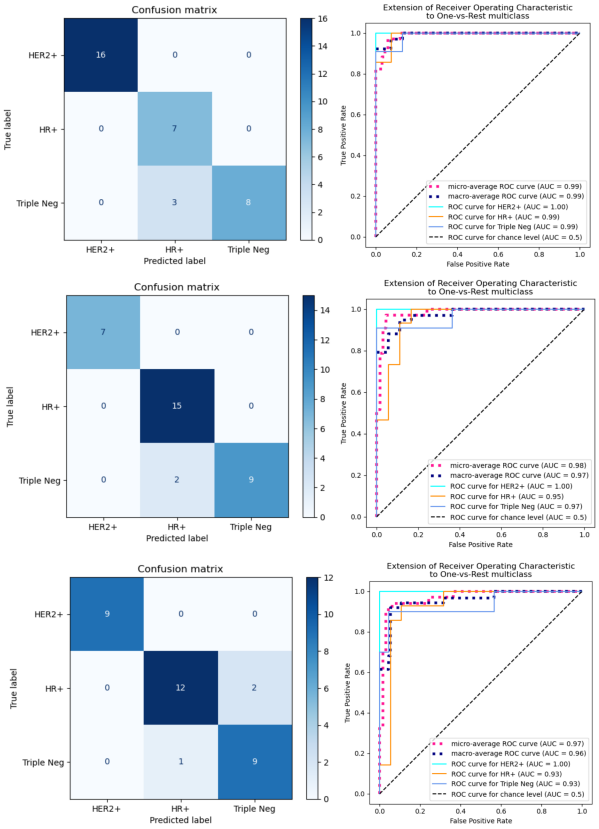

Fig. 3: *Confusion matrices and ROC curves of the optimized Random Forest classifier with 3-fold cross-validation: regardless of which subset of the data the classifier trained with, the performance remains high*

However, there is still a chance that this model is overfitted. Although we applied nested cross-validation to ensure the model selection does not include test data and also to get more realistic evaluation metrics, the results are too good to be true. The number of samples we used to train and optimize our model is low, and feature selection was performed on a high number of variables, from which we only selected 17, which might also result in some level of overfitting. Fig. 3 shows

the three confusion matrices with their corresponding ROC plot of the best Random Forest classifier that was generated per cross-validation fold. It shows that regardless of which part of the data is seen by the model for training, it performs similarly on the remaining test set.

3.3 Biomarker Selection

According to our feature selection methods, the most promising biomarker region is 17 (35076296:3528208) which was selected as the most important feature by both methods. According to the BasepairToGeneMap file this region includes the following genes: PNMT (Phenylethanolamine N-methyltransferase), PERLD1 (PER1-like domain-containing protein 1), ERBB2 (Receptor tyrosine-protein kinase erbB-2 Precursor), C17orf37 (Uncharacterized protein C17orf37), GRB7 (Growth factor receptor-bound protein 7), IKZF3 (Zinc finger protein Aiolos), AC079199.2 (ribosomal protein L39) and partially the ZPBP2 (Zona pellucida-binding protein 2 Precursor) gene. As it was expected many of these proteins are involved in signalling, and we know that the defect of cell signalling is one of the hallmarks of cancer. In addition, some of these genes were previously associated with breast cancer [4] [7], and for example, GRB7 is directly linked with HER2+ breast cancer subtype [8].

If we look at the 100 sample we trained our model with, we see that when there is an amplification at this region the subtype is HER2+, and when a sample is HER2+ subtype it always has an amplification at this region. In contrast, no clear relation can be observed based on this 100 samples for HR+ and Triple Negative samples - they can have normal copy number, gain or even loss at this biomarker region, but never amplification. So, using only this region as biomarker we could expect 100% accuracy for the HER2+ subtype, and a 50-50 chance for the other subtypes, resulting in around 60-70% accuracy if we assume that the three subtypes can be found in the population with the same probability.

4 Discussion

In this project we showed that aCGH profile can be used to predict breast cancer subtype. We have developed and optimized 4 different kinds of classifiers using two feature selection methods, and used five different performance metrics to compare the model's performance, and select the classifier that predict the breast cancer subtypes with the highest performance.

We found that all model performed well after feature selection, however, the optimized Random Forest outperformed all the other model. We observed during nested cross-validation that the performance of the model depends on the set of samples it uses for training, therefore there is a high chance that we are overestimating the performance of the Random Forest model. This effect could be decrease if the model is trained on more samples. In addition, hyperparameter optimization did not provide us with a clear answer regarding which model has the best performance, as during hyperparameter optimization multiple models were selected based on the subset of the data - therefore our final model might be a high performing one for some samples, but might have poor performance for other samples. A possible solution for this issue could be to use multiple models for breast cancer subtype prediction and look for a majority vote, or to ensemble different models.

In this project we applied feature selection to reduce noise and decrease overfitting, and it also helped us to identify important features, and therefore, chromosomal regions that could be potential biomarkers. We applied two feature selection methods, and all found the 17 (35076296:3528208) chromosomal region the most important feature. It contains genes that were previously associated with breast cancer, or even specifically with one subtypes. However, further investigation needed to understand how this region could be used as a single biomarkers, especially because looking at the 100 samples we used, this region can identify only HER2+, but can distinguish poorly between the other two subtypes.

Although, further improvement and model evaluation are needed as the current model is could be overfitted, we have proposed and proven the possibility of an aCGH-based methodology for breast cancer classification that can facilitate diagnosis and treatment decisions. In addition, further investigation regarding the found biomarker region and its genes can help us further understand how breast cancer is developed in the human organism.

5 Author Contribution

Table 2: *Author contribution: EB - Emmanouil Boutetos, LK - Liliana Kadar, MZ - Molin Zheng, RZ - Robert Zomerdijk*

Task	Contributor
Introduction	LK
Data transformation	RZ
Baseline model	LK
Feature selection	LK, MZ
PLS-DA	MZ
Random Forest	LK
Neural Network	LK, RZ
Long Short-Term Memory	EB
Biomarker selection	EB, LK
Discussion	LK
Report	LK
Report review	EB, MZ, RZ
Skeleton script	EB, LK, RZ

6 References

References

[1] Erik H van Beers and Petra M Nederlof. "Array-CGH and breast cancer". In: *Breast Cancer Research* 8 (2006), p. 210. DOI: 10.1186/bcr1510. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1557735/.

[2] Anna Bergamaschi et al. "Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer". In: *Genes, Chromosomes and Cancer* 45.11 (2006), pp. 1033–1040. DOI: https://doi.org/10.1002/gcc.20366. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gcc.20366. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/gcc.20366.

[3] Xi Chen and Hemant Ishwaran. "Random forests for genomic data analysis". In: *Genomics* 99.6 (2012), pp. 323–329. ISSN: 0888-7543. DOI: https://doi.org/10.1016/j.ygeno.2012.04.003. URL: https://www.sciencedirect.com/science/article/pii/S0888754312000626.

[4] Marlene A. Dressman et al. "Gene Expression Profiling Detects Gene Amplification and Differentiates Tumor Types in Breast Cancer". In: *Cancer Research* 63.9 (May 2003), pp. 2194–2199. ISSN: 0008-5472. eprint: https://aacrjournals.org/cancerres/article-pdf/63/9/2194/2514092/ch0903002194.pdf.

[5] Benjamin A Goldstein, Eric C Polley, and Farren B S Briggs. "Random forests for genetic association studies". en. In: *Stat. Appl. Genet. Mol. Biol.* 10.1 (July 2011), p. 32.

[6] Göran Jönsson et al. "Distinct genomic profiles in hereditary breast tumors identified by array-based comparative genomic hybridization". In: *Cancer Research* 65 (Sept. 2005), pp. 7612–7621. DOI: 10.1158/0008-5472.CAN-05-0570. URL: https://pubmed.ncbi.nlm.nih.gov/16140926/ (visited on 03/05/2021).

[7] Seung Eun Lee et al. "Phenylethanolamine N-methyltransferase downregulation is associated with malignant pheochromocytoma/paraganglioma". In: *Oncotarget* 7.17 (Mar. 2016), pp. 24141–24153. DOI: https://doi.org/10.18632/oncotarget.8234.

[8] Y. Nadler et al. "Growth factor receptor-bound protein-7 (Grb7) as a prognostic marker and therapeutic target in breast cancer". In: *Annals of Oncology* 21.3 (Mar. 2010), pp. 466–473. ISSN: 0923-7534. DOI: 10.1093/annonc/mdp346. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2826097/ (visited on 05/15/2023).

[9] A. B. Olshen et al. "Circular binary segmentation for the analysis of array-based DNA copy number data". In: *Biostatistics* 5 (Oct. 2004), pp. 557–572. DOI: 10.1093/biostatistics/kxh008.

[10] A. A. Onitilo et al. "Breast Cancer Subtypes Based on ER/PR and Her2 Expression: Comparison of Clinicopathologic Features and Survival". In: *Clinical Medicine & Research* 7 (June 2009), pp. 4–13. DOI: 10.3121/cmr.2009.825.

[11] Getiria Onsongo et al. "CNV-RF Is a Random Forest–Based Copy Number Variation Detection Method Using Next-Generation Sequencing". In: *The Journal of Molecular Diagnostics* 18.6 (2016), pp. 872–881. ISSN: 1525-1578. DOI: https://doi.org/10.1016/j.jmoldx.2016.07.001. URL: https://www.sciencedirect.com/science/article/pii/S1525157816301313.

[12] Erasmo Orrantia-Borunda et al. *Breast Cancer*. Exon Publications, Aug. 2022. URL: https://www.ncbi.nlm.nih.gov/books/NBK583808/ (visited on 04/15/2023).

[13] Lin Shi et al. "Variable selection and validation in multivariate modelling". In: *Bioinformatics* 35.6 (2019), pp. 972–980.

[14] Andrew H Sims et al. "Origins of breast cancer subtypes and therapeutic implications". In: *Nature Clinical Practice Oncology* 4 (Sept. 2007), pp. 516–525. DOI: 10.1038/ncponc0908.

[15] Mark A. van de Wiel et al. "CGHcall: calling aberrations for array CGH tumor profiles". In: *Bioinformatics* 23 (Jan. 2007), pp. 892–894. DOI: 10.1093/bioinformatics/btm030. URL: https://academic.oup.com/bioinformatics/article/23/7/892/217759 (visited on 10/18/2021).

7 Authors' Notes

7.1 Permutation studies

Using label permutation to measure model performance was emitted from this report due to lack of time.

8 Appendix

8.1 Selected features

Table 3 and 4 list the 17 features that were selected by the MUVR and Gradient Boost Classifier feature selection method respectively.

Table 3: *Top 17 most important features according to the MUVR package*

Chromosome	Start	End	Feature Rank
17	35076296	35282086	1.0000
22	41307174	41912419	563.9056
17	41062669	41447005	714.7934
12	85450052	85962613	968.5179
22	26999547	27819338	1270.4490
6	40747924	42826227	1522.3878
5	69274433	70345552	1524.3801
12	97852156	98629015	1524.4770
6	135286400	135498058	1624.8571
3	143686906	144718154	1624.9056
17	42161364	42296514	1674.3546
3	196908262	196937230	1725.0536
12	84542006	85443011	1726.0765
17	40859120	40926154	1826.4821
5	150148732	150201145	1826.7551
6	135204537	135259324	1877.4413
16	69391442	69397102	1927.7679

Table 4: *Top 17 most important features when using Gradient Boosting Classifier for feature selection*

Chromosome	Start	End	Importance Score
17	35076296	35282086	0.492572
22	26999547	27819338	0.066298
22	41307174	41912419	0.056814
5	150148732	150201145	0.056714
14	105602756	105630089	0.035010
4	162347852	163724527	0.034772
1	30716764	30984527	0.031881
15	32447707	32511475	0.031412
15	27009381	32443495	0.026263
8	36512939	36612488	0.016059
1	149579637	149582884	0.013893
11	82288127	82586347	0.013443
8	36426253	36481039	0.011146
11	82599955	82867442	0.009866
1	193480150	193527058	0.005450
13	27343854	27397290	0.005362
6	655129	1291357	0.004506

8.2 PLS-DA

Fig. 4 shows the score plot of the PLSDA, coloring the data points according to breast cancer subtypes. Table 5 list the top 17 features that were selected by PLS-DA s the most important ones.

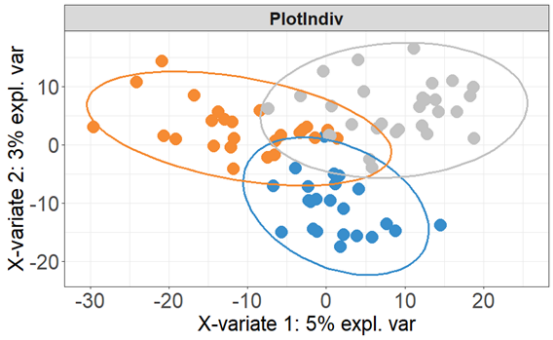

Fig. 4: *Score plot of PLSDA*

Table 5: *Top 17 most important features when using PLS-DA for feature selection*

Chromosome	Start	End	VIP score
17	35076296	35282086	2.845969451
17	42161364	42296514	2.762316684
17	35076296	35282086	2.7503681094
17	40859120	40926154	2.746213168
22	41307174	41912419	2.716214199
17	38784700	39114890	2.659289698
12	85450052	85962613	2.593881532
12	97852156	98629015	2.563022152
16	69391442	69397102	2.515372032
12	84542006	85443011	2.488515653
16	32378126	32748349	2.488515653
17	1456417	1524691	2.480597887
4	162347852	163724527	2.473662615
11	82599955	82867442	2.456916347
1	3810936	6606212	2.44482736
1	27650970	27755668	2.44482736
3	143686906	144718154	2.435034865

8.3 LSTM Validation Loss

Fig. 5 illustrates the training and validation loss curves, which was used to monitor the performance of the LSTM model during training, helping to prevent overfitting and to decide the number of the epochs for our model.

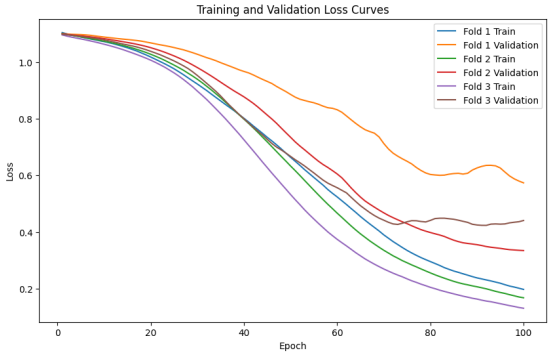

Fig. 5: *Loss and Validation Loss for the 3 outer folds during fitting*